Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Kratos: an auditable agentic workflow for automated calibration and population estimation of target-mediated drug disposition models

Devam A Desai

desaidevam@gmail.com

University of Florida

Rodrigo Cristofoletti

University of Florida

Research Article

Keywords: Target-mediated drug disposition, Monoclonal antibodies, Agentic workflow, Automation, nlmixr2, Reproducibility

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11496689/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Kratos: an auditable agentic workflow for automated calibration and population estimation of target-mediated drug disposition models

Devam A. Desai and Rodrigo Cristofoletti

Corresponding author: Devam A. Desai, desaidevam@gmail.com, ORCID 0009-0006-8880-8718

Abstract

Building target-mediated drug disposition (TMDD) models involves many repetitive, error-prone steps. We developed Kratos, a proof-of-concept auditable agentic workflow in which a deterministic controller dispatches specialised agents for data validation, model specification, bounded multistart calibration with an adaptive rules planner, nlmixr2 FOCEi population estimation, diagnostics and reporting, and records every decision and numerical attempt. All analyses used the rules planner; no large language model (LLM) was called. Kratos was verified by cross-engine comparison, an automated test suite and recovery of known parameters from simulated data, and was applied to 143 published mean concentration–time profiles of 36 human monoclonal antibodies, split by molecule into two parts that were analysed independently and used to predict each other. On simulated data all 6 estimated typical parameters were recovered within 9% of truth, Python and R predictions agreed to a relative difference of 1.2×10^{-6} or less,

and 54/54 automated tests passed. For the antibody data, linear disposition and absorption parameters were consistent across parts (clearance 0.188 and 0.268 L/day; bioavailability 0.71 and 0.84), whereas target-mediated parameters were not identifiable; the Part B covariance step failed and was flagged. Each part's model predicted the other part's unseen molecules with median fold errors of 1.29 and 1.35. The workflow ran from raw data to report without analyst intervention and kept numerical failures visible. Accuracy and reproducibility were supported; an efficiency advantage of adaptive over fixed multistart planning was not demonstrated.

Keywords: Target-mediated drug disposition; Monoclonal antibodies; Agentic workflow; Automation; nlmixr2; Reproducibility

Introduction

Target-mediated drug disposition (TMDD) describes the nonlinear pharmacokinetics (PK) that arise when a drug binds its pharmacological target with high affinity and the drug–target complex is eliminated [1]. TMDD is common among therapeutic monoclonal antibodies (mAbs), whose disposition combines linear, FcRn-protected catabolism with saturable target-mediated elimination [2–5]. The full TMDD model and its quasi-equilibrium, quasi-steady-state and Michaelis–Menten approximations are well characterised [6–8], as are their characteristic profiles and the conditions under which their parameters can be identified [9].

Fitting these models in practice is laborious. An analyst must reconcile units and molecular weights, encode dosing events, choose initial estimates for a stiff ordinary differential equation (ODE) system with up to eleven structural parameters, run multistart optimisation, carry candidate solutions into nonlinear mixed-effects estimation, review convergence, covariance, shrinkage and residual diagnostics, and document every decision [10–12]. Each step is a potential source of error, and quality-control guidance stresses traceability from source data to reported result [13, 14].

Automation of pharmacometric model development has so far focused mainly on searching model structure. Examples include genetic-algorithm and machine-learning model selection [15–17] and fully automatic population PK model development [18]. More recently, large language models (LLMs) and "agentic" systems, in which a model plans and invokes tools [19], have been explored for writing and interpreting NONMEM code and for orchestrating analyses [20–24]. PKGPT uses an LLM agent to generate, execute and refine NONMEM population PK models, with evaluation against human expert models [24]. GRASP, reported as a preprint, uses graph-based agents to construct and modify mechanistic quantitative systems pharmacology (QSP) models, including TMDD extensions, through natural-language interaction and MATLAB/SimBiology execution [25]. These developments motivate explicit verification of both the numerical analysis and the decisions made during automated workflows.

We developed Kratos as a proof-of-concept workflow for auditable estimation with a fixed, approved full TMDD model. Its contribution is the integration of bounded multistart calibration, nonlinear mixed-effects

estimation, cross-engine numerical verification and failure reporting within a reproducible path from input data to report. A deterministic controller validates every proposed action and records every numerical attempt. The present evaluation uses a deterministic rules planner; the optional LLM planner is outside its empirical scope. Our central hypothesis was that such a constrained, auditable workflow can reduce the analyst effort needed to calibrate and estimate TMDD models while preserving (H1) estimation accuracy, (H2) numerical reliability and (H3) reproducibility, and that (H4) its adaptive planner reaches acceptable fits more efficiently than a non-adaptive multistart baseline. Here we describe the workflow, report its verification on simulated data, apply it to a pooled set of published human mAb mean PK profiles with an out-of-sample test, and evaluate each sub-hypothesis, including those that were not supported.

Methods

Workflow architecture

Kratos is implemented in Python with an R back end (Fig. 1). A deterministic controller (the orchestrator) executes a fixed sequence of stages: data validation, raw-data graphics, model specification, calibration, population estimation, diagnostics and reporting. It dispatches each stage to a specialised agent. Here, an agent is a task-specific software component operating through explicit input and output contracts; agentic describes bounded coordination and feedback-driven action selection. In the evaluated implementation, these components use programmed rules and numerical tools. Agents communicate only through typed action and result objects that are validated against JSON schemas before execution; for example, a proposed calibration start is rejected if it lies outside the declared bounds or modifies a fixed parameter. Budgets (number of attempts, function evaluations, wall time) and stopping rules are enforced by the controller, not by the agents. An optional LLM planner can be enabled for calibration. It receives only an approved diagnostic summary (no subject-level records), must return schema-valid JSON, and falls back to the deterministic rules planner on any invalid or missing response. All analyses in this report used the deterministic rules planner, and no LLM call was made. The evaluation therefore concerns rules-based orchestration and adaptive heuristics, without evaluating LLM reasoning or autonomous model generation.

Each run is stored in its own directory. The directory contains the original inputs and their SHA-256 hashes, the normalised dataset and subject-ID map, a configuration snapshot, all random seeds, a hash of the source code, an append-only audit log, the complete calibration history, the nlmixr2 fit object, tables, every figure together with the data used to draw it, a self-contained report and a checksummed manifest. Figures and reports can be regenerated from saved results without refitting. Runs can be resumed or cancelled without losing completed work.

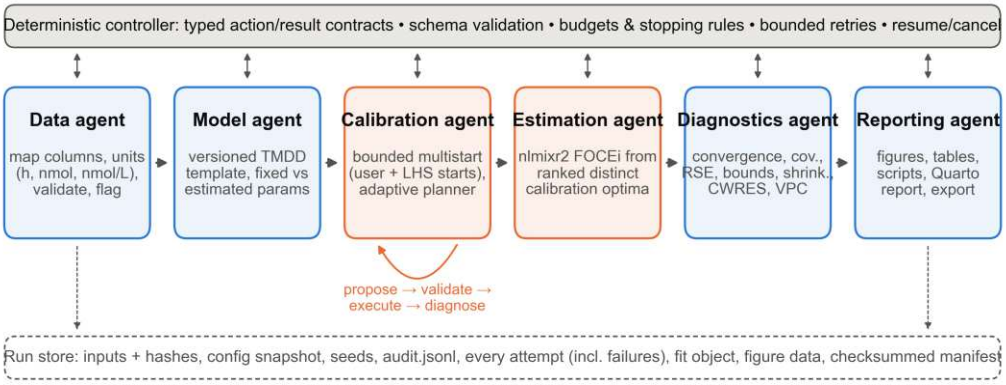


Fig. *1 Architecture of Kratos. A deterministic controller dispatches six agents through typed, schema-validated action and result contracts and enforces budgets, stopping rules and bounded retries. The calibration agent iterates a propose–validate–execute–diagnose loop with an adaptive rules planner. The reported analyses used this deterministic planner; the optional LLM planner was not invoked. All inputs, decisions, attempts (including failures), fits and figure data are written to a hashed, checksummed run store.*

Data handling

The data agent maps user columns to a canonical event schema, derives dosing events (intravenous [IV] bolus, finite infusion, repeat dosing via additional-dose records, and subcutaneous [SC] dosing into a depot), and converts all quantities to hours, nmol and nmol/L using analyst-declared units and molecular weights. Every conversion is recorded. Records that the model cannot represent faithfully are rejected with row- and field-specific messages, including steady-state shortcuts, modelled rates, unknown routes, inconsistent infusion rate and duration, censored records and non-positive concentrations. Suspicious records (for example, time inversions or unexpected rises after a single IV dose) are flagged for review and never deleted.

Structural and statistical model

All analyses used an approved, versioned template (tmdd2c_full v1.1.0): a two-compartment full TMDD model with 1:1 binding in the central compartment and a first-order SC absorption depot [1]:

$$\frac{dA_d}{dt} = -k_a A_d \qquad \text{(SC dose} \times F \text{ enters } A_d\text{)}$$
$$\frac{dA_c}{dt} = k_a A_d + \mathrm{Input}_{IV}(t) - CL\,C - Q\,(C - C_p) - V_c\,(k_{on}\,C\,R - k_{off}\,RC)$$
$$\frac{dA_p}{dt} = Q\,(C - C_p)$$
$$\frac{dR}{dt} = k_{syn} - k_{deg}\,R - k_{on}\,C\,R + k_{off}\,RC$$
$$\frac{dRC}{dt} = k_{on}\,C\,R - (k_{off} + k_{int})\,RC$$

where $C = A_c/V_c$ is the free drug concentration, $C_p = A_p/V_p$, R and RC are the free-target and complex concentrations, and $k_{syn} = k_{deg}R_0$, so that the drug-naive baseline is $R(0) = R_0$. Observations were

mapped to free drug C. Between-subject variability was log-normal, $\theta_i = \theta_{pop}\exp(\eta_i)$ with $\eta \sim N(0, \Omega)$ and Ω diagonal. Residual error was log-normal, $\log y = \log f + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2)$. Structural changes to the template require a new version and an analyst decision that is recorded in the configuration.

Calibration

Pooled calibration (one parameter set for all individuals, no random effects) supplied starting values for population estimation. The objective was the sum of squared log residuals, which corresponds to the log-normal residual model. Each attempt was minimised with a bounded trust-region reflective least-squares algorithm [26, 27] in log-transformed (logit for F) parameter space. The first round consisted of the analyst's initial values plus Latin hypercube points [28] spanning the declared bounds.

The adaptive rules planner then proposed further starts from the accumulated history:

- if no attempt had succeeded, failed starts were moved halfway toward the centre of the bounds;
- otherwise, a parameter that ended near a bound (within 2% of its log- or logit-transformed range) was pulled back toward the centre to test whether the boundary solution persisted;
- local perturbations were drawn around the best fit, with a spread that shrank as successful attempts accumulated and that excluded neighbourhoods of failed starts;
- one unused Latin hypercube point was added per round for global exploration.

The non-adaptive baseline (fixed planner) consumed only Latin hypercube points. Calibration stopped when the attempt, evaluation or time budget was exhausted, when the best objective had been reproduced within 0.1% by a prespecified number of independent starts, or when it had not improved by more than 0.1% over a window of successful attempts. Every attempt was recorded with its start values, result, number of evaluations, run time and messages.

Population estimation

Up to three distinct, ranked calibration optima were translated into nlmixr2 models [29], using rxode2 for the ODE solution [30]. Thetas were placed on the log or logit scale within the declared bounds, and fixed parameters used fix(). The models were fitted by first-order conditional estimation with interaction (FOCEi) [31], with the sandwich covariance matrix (r,s) and the LSODA solver [32] (relative and absolute tolerances 10^{-8} and 10^{-10}). Candidates were tried in rank order with bounded retries, and the first fit with a finite objective function value (OFV) was accepted. Wald 95% confidence intervals (CIs) were computed on the estimation scale and back-transformed. Relative standard errors (RSEs) were obtained on the parameter scale by the delta method.

Diagnostics

The diagnostics agent assessed:

- optimiser convergence and the covariance step;

- RSEs and the proximity of estimates to their bounds;
- parameter correlations and η-shrinkage [11];
- conditional weighted residuals (CWRES) against time and population prediction [33];
- a visual predictive check (VPC) with 200 replicates that preserved each individual's dose history, stratified by route (mAb data) or design group (simulated data), dose-normalised for the mAb data, and binned at quantiles of observation time [34].

Review flags were raised for RSE > 50%, shrinkage > 30%, |correlation| > 0.95, an estimate within 2% of the transformed parameter range from a bound, |Spearman ρ| > 0.3 between CWRES and time or prediction, or more than 20% of observed percentile points outside the simulated 95% CIs. These thresholds are heuristic prompts for scientific review, not acceptance criteria. Kratos does not declare a model valid.

Verification studies

Cross-engine agreement. The Python engine (SciPy BDF solver) and the R engine (rxode2) were compared on identical events and parameters. Four cases were run: typical parameters under pre- and post-dose conventions, a high-capacity target, and the 36 individual parameter sets of the simulated dataset. The prespecified criterion was |difference| ≤ 10^{-6} nmol/L + 10^{-4} × |value|.

Automated tests. A pytest suite covered:

- hand-checkable dosing and mass-balance cases;
- rejection of invalid inputs;
- planner contracts and optional LLM-interface guardrails (using mocked responses, without a live LLM);
- cancellation and resume;
- exclusion of credentials from exports;
- seed reproducibility;
- regeneration of all outputs without refitting.

Parameter recovery. 36 subjects in six design groups were simulated from known typical parameters:

- IV bolus 0.1, 0.3 and 1 mg/kg;
- 1-h infusions of 3 and 10 mg/kg;
- 1 mg/kg every 2 weeks for three doses, with pre-dose troughs.

The simulation used log-normal between-subject variability on CL, V_c and R_0 (ω^2 = 0.09, 0.04 and 0.09) and a log-normal residual SD of 0.1. Binding and turnover constants (k_{on}, k_{off}, k_{deg}) were fixed at their true values. Samples below 0.05 nmol/L (111 samples) were not recorded, because censored-data

likelihoods were outside the scope of this version. Recovery was prespecified as successful if every estimated typical value lay within ±25% of truth. A subject-level nonparametric bootstrap [35] refitted FOCEi to resampled datasets with dose histories preserved.

Reproducibility. All three analyses reported here were run twice, as independent end-to-end executions from the same data, configuration and seeds, and the saved estimates were compared.

Application: human mAb mean PK profiles

The application reused the clinical mAb PK database assembled by Desai and Cristofoletti for their low-dimensional neural ordinary differential equation (NODE) study [36]. The database comprised 1,423 concentration observations from 143 published mean concentration–time profiles of 36 therapeutic mAbs studied in humans after IV or SC administration. Each profile represented one molecule, dose and route. The underlying clinical publications and regulatory review documents are identified in Table 1 and the reference list of the companion manuscript [36].

The database-building strategy followed the inter-antibody PK workflow of Kumar et al. [37]. Profiles were selected from publicly available clinical pharmacology publications and regulatory reviews when they provided sufficient concentration–time information to characterise disposition and, for SC administration, absorption. Trough-only profiles were excluded. For multiple-dose studies, data from the first dosing cycle were used. Published PK profiles were digitised using WebPlotDigitizer (version 5.2) [38] and compiled into a harmonised dataset. Trastuzumab was not included [36].

For the Kratos analysis, the full database was repartitioned independently of the NODE study's 31-antibody training and five-antibody test split. Profiles were assigned to two parts by a seeded (seed 20260929), molecule-grouped, balanced split, so that all doses and routes of an antibody fell into the same part and the two parts shared no molecule. Part A comprised 72 profiles (59 IV, 13 SC) of 19 molecules and Part B comprised 71 profiles (58 IV, 13 SC) of 17 molecules (Online Resource 1).

For the Kratos reanalysis, the analyst specified the following data-harmonisation and observation-model settings in the configuration:

- time in hours, amounts in µg and concentrations in µg/mL;
- a molecular weight of 150 kDa for every antibody;
- an assay measuring free drug;
- IV doses treated as boluses;
- drug-naive baseline target levels.

These settings distinguish the harmonised input units from the common molecular weight, free-drug observation model, bolus approximation for IV administration and baseline-target assumptions used in

Kratos. The latter assumptions were not established as molecule- and study-specific properties by the companion dataset description [36].

Four source rows that carried both a dose and a time-zero concentration were split into a dose record and an immediately post-dose observation. Because the data contain no target measurements and the targets differ between molecules, k_{on} (1 L·nmol^{-1}·h^{-1}), k_{off} (0.1 h^{-1}) and k_{deg} (0.1 h^{-1}) were fixed at generic values. CL, V_c, Q, V_p, k_{int}, R_0, k_a and F were estimated, with random effects on CL, V_c and R_0. Because each "individual" is a mean profile, Ω describes between-profile variability that combines differences between molecules, studies, doses, routes, study populations, assays and digitisation, rather than between-patient variability [36]. Standard errors do not reflect the number of patients behind each mean. Each part was analysed independently with identical configurations except for the input file.

Out-of-part prediction. The typical values, Ω and residual SD estimated from one part were used to simulate the other part's dosing design (200 replicates). Population predictions were compared with observations using the median fold error (MFE; the exponential of the median absolute log error [39]), the fraction of observations within two-fold, and the coverage of the 90% prediction interval (PI).

Planner evaluation (H4)

The adaptive rules planner and the fixed Latin hypercube baseline were compared on replicated simulated datasets (same design, four subjects per design group) generated with new seeds (five replicates in each setting). In each replicate, both arms used the same data, model, bounds, evaluation budget, stopping rules and calibration seed. Two settings were studied:

- *Default:* the analyst's initial values plus three Latin hypercube points in the first round, with early stopping.
- *Stress:* a poor initial guess far from the truth (R_0 = 90 vs. 2 nmol/L; k_{int} = 4 vs. 0.05 h^{-1}), a smaller first round (one Latin hypercube point) and two attempts per round, which made failed attempts more likely.

The LLM-assisted arm was not run. This comparison evaluates deterministic rules-based adaptation against fixed multistart planning.

Software

Analyses used:

- Python 3.12.0 with NumPy 1.26.4 [40], SciPy 1.14.1 [27], pandas 2.2.2 and Matplotlib 3.9.1;
- R 4.5.1 [41] with rxode2 5.1.6, nlmixr2 4.0.1 and nlmixr2est 4.1.0;
- Quarto 1.10.18 for reporting [42].

Computations ran on a 12-core Windows 10 workstation. Numerical analysis results, including those reported in the text, are read from saved run artifacts by a script distributed with the code; dataset provenance and acquisition methods are reported from the companion manuscript [36].

Results

Platform verification (H1–H3)

All 1362 cross-engine comparisons in 4 cases met the prespecified criterion, with a maximum relative difference of 1.2×10^{-6} between the Python and R predictions. The automated suite passed 54 of 54 tests (0 skipped), including LLM-interface guardrail tests with mocked responses. These tests do not establish the performance of a live LLM planner.

On the simulated dataset (36 subjects, 387 observations), calibration stopped after 6 attempts (4 succeeded). FOCEi converged (relative convergence (4)) with a successful covariance step. All 6 estimated typical values were within 8.6% of truth, meeting the ±25% criterion (Table 1, Fig. 2a). The true value lay inside the Wald 95% CI for 5 of 6 parameters; the exception was Q (estimate 0.0326 vs. true 0.03; error 8.6%), whose narrow interval excluded the truth. Variance components were estimated less accurately: the estimated ω^2 was 0.034 for CL (true 0.09), 0.033 for V_c (true 0.04) and 0.073 for R_0 (true 0.09), and the residual SD was 0.099 (true 0.1). In the bootstrap, 5 of 8 replicates converged; the remainder ended with ‘false convergence’ and were excluded rather than repaired. The bootstrap percentile intervals agreed with the Wald intervals (Fig. 2a). The VPC reproduced the observed percentiles in all six design groups (4% of percentile points outside the simulated intervals; Fig. 2b), and the fit raised 0 critical and 0 warning flags.

The repeated, independent executions of all three analyses reproduced the calibration histories exactly, and all parameter estimates and OFVs were identical to full numerical precision.

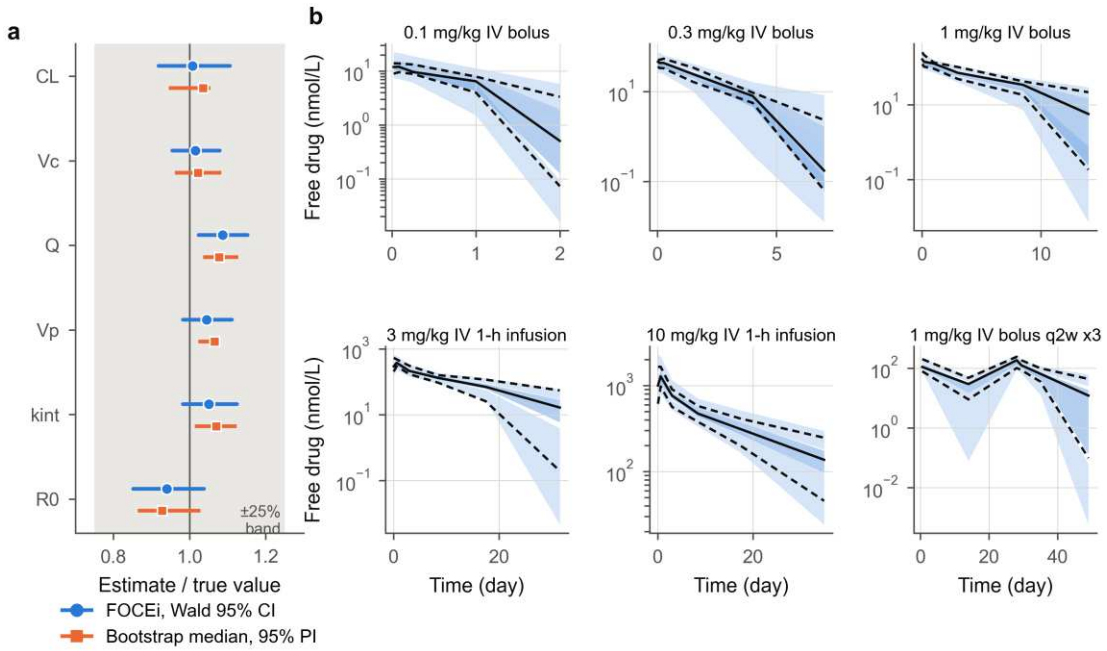


Fig. *2 Verification on simulated data with known parameters.* ***a*** *Ratio of estimate to true value for the estimated typical parameters (circles, FOCEi with Wald 95% CI; squares, bootstrap median with percentile 95% interval); the shaded band marks the prespecified ±25% recovery criterion.* ***b*** *Visual predictive check by design group (200 replicates): observed 5th, 50th and 95th percentiles (lines) and simulated 95% CIs of the same percentiles (shaded).*

Table *1 Recovery of known typical parameters from the simulated dataset (FOCEi). CI, Wald 95% confidence interval; bootstrap, percentile 95% interval from successful bootstrap replicates.*

Parameter	Unit	True	Estimate	RSE (%)	95% CI	Bootstrap 95% interval	Error (%)
CL	L/h	0.01	0.0101	4.7	0.00919–0.0111	0.00951–0.0105	0.8
Vc	L	3	3.04	3.1	2.86–3.24	2.9–3.23	1.5
Q	L/h	0.03	0.0326	3	0.0307–0.0346	0.0312–0.0337	8.6
Vp	L	2.5	2.61	3.2	2.46–2.78	2.57–2.68	4.5
kint	1/h	0.05	0.0525	3.5	0.0491–0.0562	0.0509–0.0559	5.1
R0	nmol/L	2	1.88	5	1.7–2.08	1.74–2.05	-6

Application to the mAb dataset

Both parts passed validation; only review flags were raised, and no record was deleted. The flags covered combined dose and observation rows (split), time inversions from digitisation (sorted, with the original preserved), one IV profile with a late rise in Part A, and the aggregate-data notice. Part A contained 744 observations (median 10.5 per profile; doses 0.7–8,400 mg) and Part B contained 679 (median 9; 2.1–

10,800 mg). After dose-normalisation, the profiles of the two parts overlapped, with biphasic IV decline and SC peaks within the first weeks (Fig. 3).

In Part A, calibration ran 8 attempts (8 succeeded) and stopped on numerical success; the pooled optima formed 1 distinct solution (Fig. 7a). In Part B, calibration ran 8 attempts (7 succeeded) and stopped on plateau; the pooled optima formed 3 distinct solutions (Fig. 7a). Both pooled optima placed most elimination on the linear pathway (R_0 0.018 and 0.28 nmol/L), which is expected when molecules with different targets share one parameter set.

FOCEi converged in both parts (relative convergence (4); relative convergence (4)). The Part A covariance step succeeded and the Part B covariance step failed because k_{int} converged to its lower bound (0.0001 h^{-1}). Kratos reported this as a critical review flag and did not suppress or re-run it. Estimation took 63 and 34 min, respectively. Table 2 summarises the estimates.

The linear disposition and absorption parameters were well determined and consistent between the independently analysed parts. Typical CL was 0.188 L/day in Part A (95% CI 0.175–0.202) and 0.268 L/day in Part B. V_c was 3.08 and 3.50 L, and V_{ss} was 5.7 and 5.8 L. The terminal half-lives implied by the linear two-compartment parameters were 22.8 and 16.5 days. The absorption rate constant k_a was 0.209 and 0.304 day^{-1} and bioavailability F was 0.71 and 0.84. These values are close to published typical values for IgG mAbs: a pooled first-in-human analysis of four mAbs reported CL 0.200 L/day, V_c 3.61 L, V_p 2.75 L, k_a 0.282 day^{-1} and F 0.744 [43], and a review of population analyses reported median central and peripheral volumes of 3.1 and 2.8 L and linear clearances of 0.2–0.5 L/day [4, 5]. In contrast, the target-mediated parameters were poorly identified. In Part A, k_{int} had an RSE of 93% and R_0 an RSE of 66%, with η-shrinkage on R_0 of 86%. In Part B, k_{int} reached its bound and the R_0 shrinkage was 67%. Between-profile variability in CL was 26% CV in Part A and 68% CV in Part B.

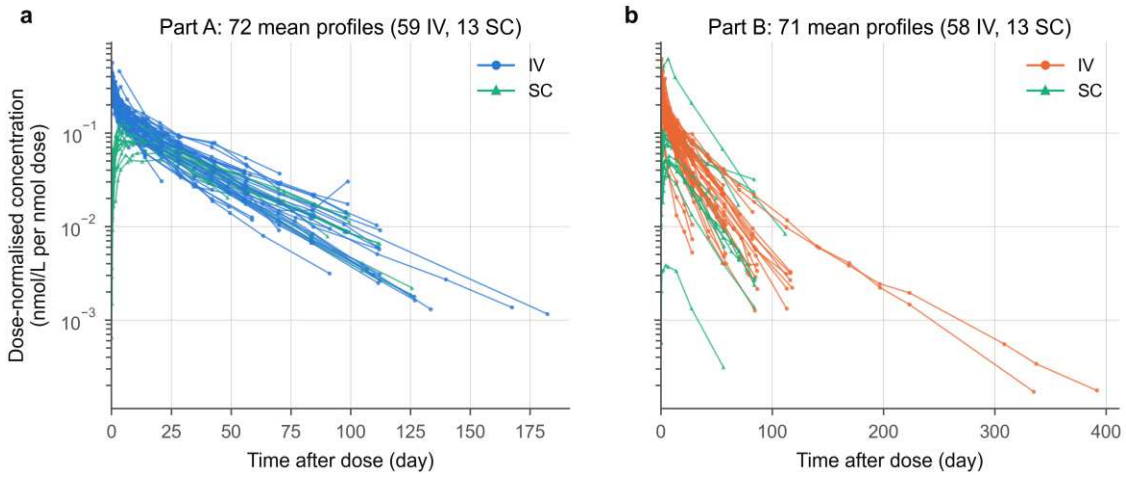


Fig. *3 Observed dose-normalised mean concentration–time profiles of the mAb dataset after IV (circles) and SC (triangles) administration for* ***a*** *Part A and* ***b*** *Part B (molecule-disjoint, analysed independently).*

***Table** 2 FOCEi estimates for the two independently analysed parts of the mAb dataset. Parameters are on the model (hourly) scale; kon, koff and kdeg were fixed. RSE, relative standard error (delta method); CI, Wald 95% confidence interval; BPV, between-profile variability (CV%); –, not available because the covariance step failed.*

Parameter	Unit	Part A estimate	Part A RSE (%)	Part A 95% CI	Part A BPV / shrinkage (%)	Part B estimate	Part B RSE (%)	Part B 95% CI	Part B BPV / shrinkage (%)
CL	L/h	0.00783	3.8	0.00727–0.00843	26 / 9	0.0112	–	–	68 / 2
Vc	L	3.08	3.3	2.88–3.28	28 / 9	3.5	–	–	69 / 2
Q	L/h	0.0214	14.5	0.0161–0.0285	–	0.0184	–	–	–
Vp	L	2.63	5.0	2.39–2.91	–	2.26	–	–	–
ka	1/h	0.00869	11.3	0.00696–0.0108	–	0.0127	–	–	–
F	–	0.705	5.9	0.618–0.78	–	0.835	–	–	–
kint	1/h	0.000396	93.0	6.4e-05–0.00245	–	0.0001	–	–	–
R0	nmol/L	0.0058	65.9	0.00159–0.0211	162 / 86	0.0382	–	–	119 / 67
kon	L/(nmol·h)	1 (fixed)	–	–	–	1 (fixed)	–	–	–
koff	1/h	0.1 (fixed)	–	–	–	0.1 (fixed)	–	–	–
kdeg	1/h	0.1 (fixed)	–	–	–	0.1 (fixed)	–	–	–
Residual SD	log scale	0.197	–	–	–	0.221	–	–	–

OFV: Part A 4005.5 (744 observations, 72 profiles); Part B 4890.7 (679, 71).

The goodness-of-fit plots showed no systematic trend in CWRES over time (Spearman ρ 0.06 and 0.01) and CWRES means near zero (0.02 and -0.01; SD 0.98 and 1.02) (Fig. 4). Population predictions were more dispersed in Part B, which is consistent with its larger between-profile variability. In the VPCs (Fig. 5), 21% (Part A) and 43% (Part B) of the observed percentile points fell outside the simulated 95% CIs, which exceeds the heuristic threshold of 20%. Discrepancies concentrated in the outer percentiles and in late time bins: the pooled model reproduced the median profile but not the spread of heterogeneous molecules. Part A raised 0 critical and 4 warning flags; Part B raised 1 critical and 4 warning flags.

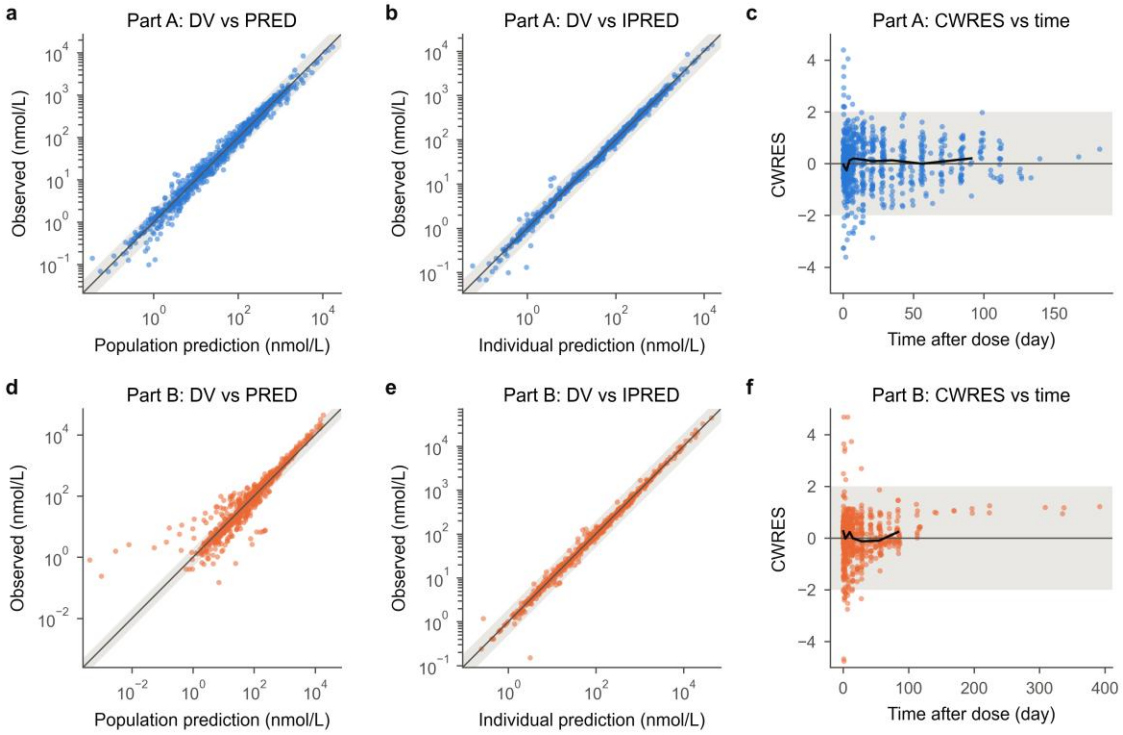


Fig. *4 Goodness of fit of the FOCEi models. Observed concentrations against population (**a**, **d**) and individual (**b**, **e**) predictions with the line of identity and a two-fold band (shaded); CWRES against time with a binned median (**c**, **f**) for Part A (top) and Part B (bottom).*

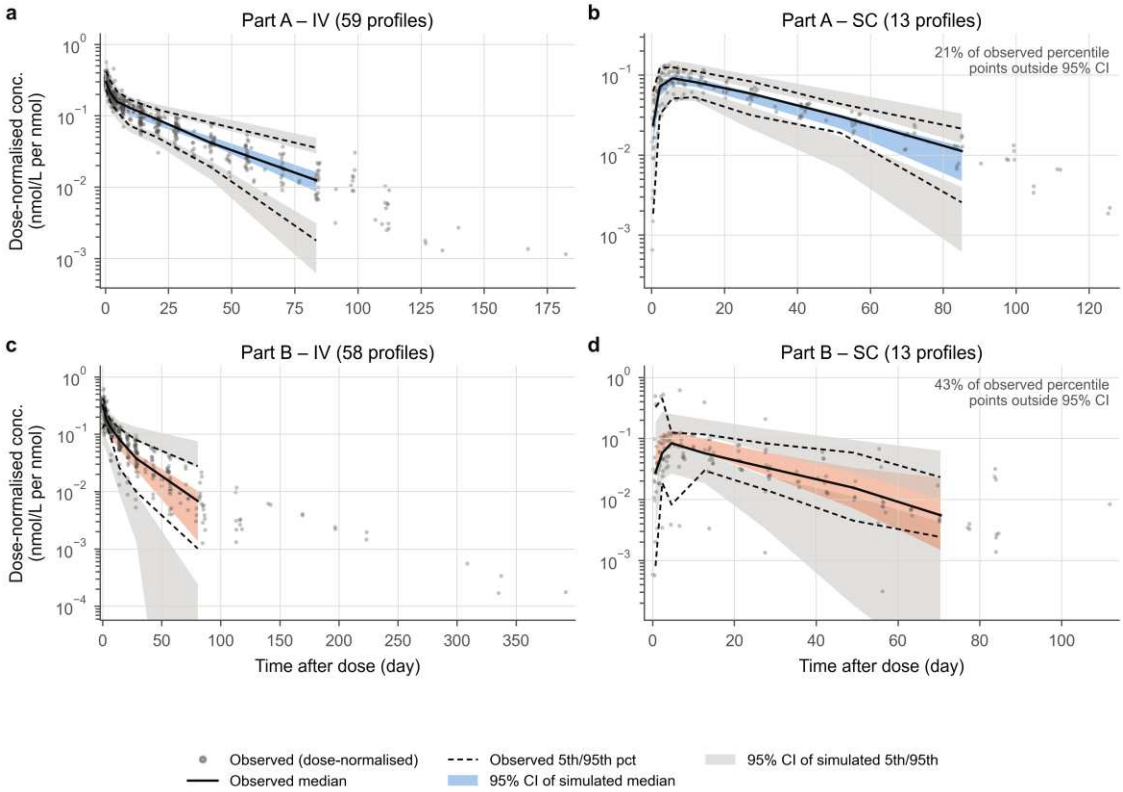


Fig. *5 Dose-normalised visual predictive checks stratified by route (200 replicates) for Part A (**a**, **b**) and Part B (**c**, **d**). Points, observations; solid and dashed lines, observed median and 5th/95th percentiles; shaded, simulated 95% CIs of the median (coloured) and of the 5th/95th percentiles (grey). Axes are limited to the observed range; in **c** the lower simulated interval extends below the plotted range in late bins.*

Out-of-part prediction

Each part's model was used to predict the other part's unseen molecules (Fig. 6, Table 3). The Part A model predicted Part B with an MFE of 1.29 (76% of observations within two-fold; 90% PI coverage 75%). The Part B model predicted Part A with an MFE of 1.35 (84% within two-fold; coverage 97%). IV profiles transferred better than SC profiles. For the Part A model applied to Part B SC profiles, the MFE was 1.80 and the 90% PI coverage was 46%. The higher coverage of the Part B model reflects its wider between-profile variance (ω^2 for CL 0.38 vs. 0.07) rather than more accurate central predictions. For reference, the Part A and Part B models fitted their own data with population-prediction MFEs of 1.21 and 1.30.

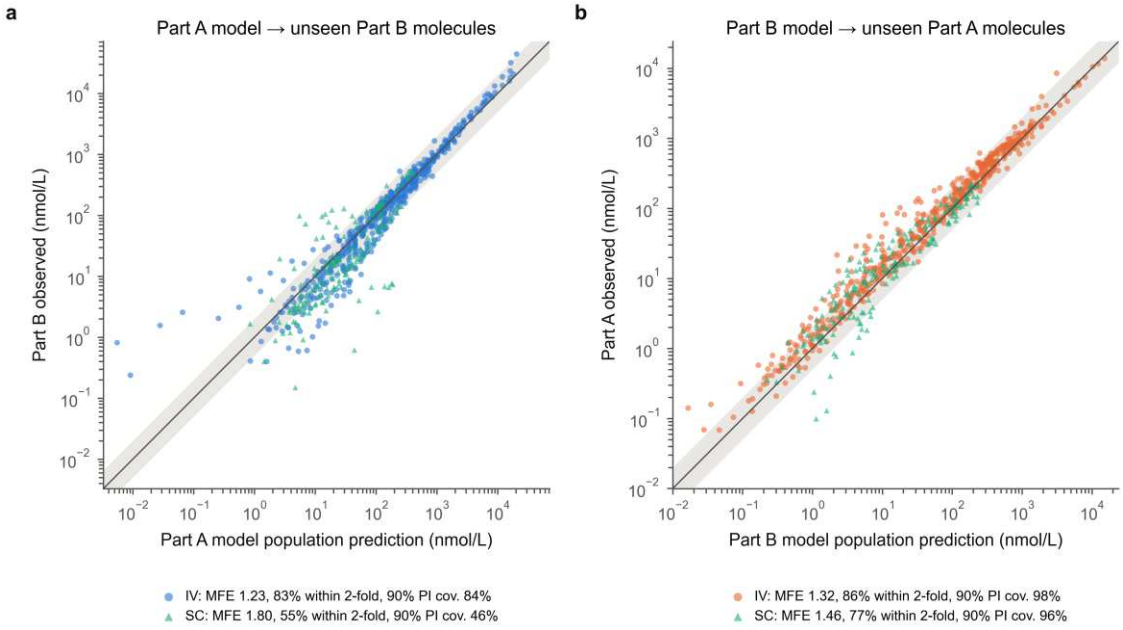


Fig. *6 Out-of-part prediction of unseen molecules.* ***a*** *Part A model predicting Part B;* ***b*** *Part B model predicting Part A. Observed concentrations against population predictions for IV (circles) and SC (triangles) profiles, with the line of identity and a two-fold band (shaded). The legend reports the median fold error (MFE), the fraction within two-fold and the 90% prediction-interval coverage by route.*

Table *3 Out-of-part prediction of unseen molecules from population parameters (typical values, Ω and residual SD; 200 simulation replicates). MFE, median fold error; PI, prediction interval.*

Model → data	Route	n obs	MFE	Median log error	Within 2-fold	Within 3-fold	90% PI coverage
Part A → Part B	All	679	1.29	-0.1	76%	89%	75%
Part A → Part B	IV	514	1.23	-0.05	83%	93%	84%
Part A → Part B	SC	165	1.8	-0.41	55%	79%	46%
Part B → Part A	All	744	1.35	0.24	84%	93%	97%
Part B → Part A	IV	561	1.32	0.26	86%	95%	98%
Part B → Part A	SC	183	1.46	0.13	77%	90%	96%

Planner evaluation (H4)

In the default setting, both arms reached the same optimum in 5/5 replicates, with identical numbers of attempts (4). This agreement was uninformative: the stopping rule was met within the first round of user and Latin hypercube starts, which both arms share, so the adaptive planner issued 0 proposals. In the stress setting, the adaptive planner made 10 proposals (local perturbations, a boundary-escape start and exploration points) across the replicates. Both arms again reached equivalent optima in 5/5 replicates with equal numbers of attempts. Wall time was lower for the adaptive arm in 5/5 replicates (mean difference -

18 s, SD 24 s; two-sided sign test p = 0.06; Fig. 7b,c), a small difference relative to the 208 s mean run time that five replicates cannot establish as systematic. No analyst intervention was required in any run.

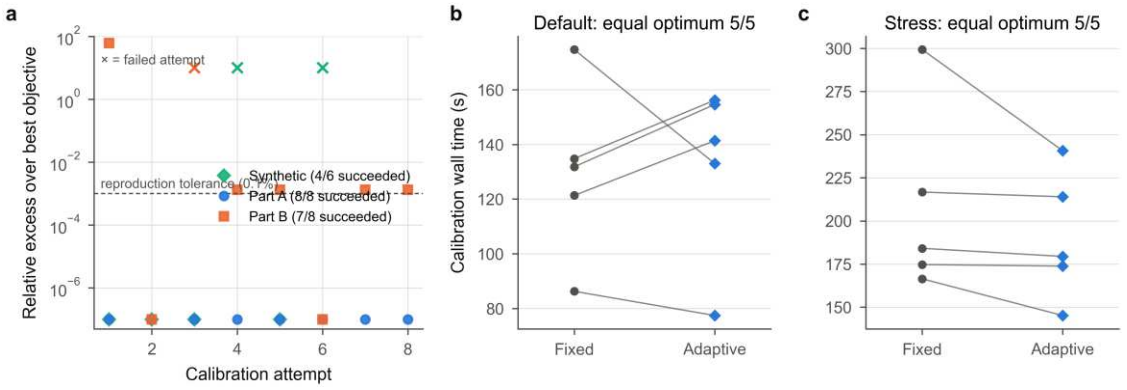


Fig. *7 Calibration behaviour and planner evaluation.* ***a*** *Relative excess of each successful calibration attempt over the best objective of its run (crosses, failed attempts; dashed line, 0.1% reproduction tolerance).* ***b***, ***c*** *Paired calibration wall time of the fixed Latin hypercube multistart and the adaptive rules planner across replicated simulated datasets in the default and stress settings.*

Discussion

Using deterministic rules-based planning, Kratos automated the full path from a raw concentration–time file to a population TMDD fit, diagnostics and a reproducible report, without analyst intervention, and made each step inspectable. The verification studies supported three of the four sub-hypotheses. Accuracy (H1) was supported for typical parameters on simulated data with known truth. Numerical reliability (H2) was supported by agreement between independent solvers well within the prespecified tolerance, by the test suite, and by the way real-data problems surfaced: in Part B, the boundary estimate of k_{int} (warning) and the failed covariance step (critical flag) were reported, not silently repaired. Reproducibility (H3) was supported by identical repeated executions and by regeneration of every output from saved artifacts. The efficiency claim (H4) was not supported, as discussed below.

The mAb application illustrates both the value and the limits of automation. Two independently analysed, molecule-disjoint subsets gave consistent linear disposition and absorption parameters, and each predicted the other's unseen molecules with median fold errors of about 1.3, which is useful for prior-informed planning of first-in-human studies [43]. The same analysis showed that k_{int} and R_0 were not identifiable from pooled, drug-only mean profiles with fixed generic binding constants. This was expected from the structure of the problem: free-drug data alone constrain target-mediated parameters only when the profiles span the nonlinear range with sufficient density, and target capacity differs between molecules [7, 9]. In this setting, a workflow that reports high RSE, high shrinkage, boundary estimates and a failed covariance step—rather than presenting a converged fit as final—behaves as a careful analyst should. The appropriate next model revision, such as a quasi-steady-state or Michaelis–Menten

reparameterisation, fixing k_{int}, or molecule-level random effects on target parameters, is a scientific decision. Kratos records such a decision as a new template version but deliberately does not make it.

Model-search frameworks such as pyDarwin and Pharmpy-based automatic model development optimise over structural and statistical model space [16, 18]. PKGPT extends agentic automation to iterative NONMEM model development with an LLM and human-expert benchmarks [24]. GRASP demonstrates graph-mediated construction and modification of mechanistic QSP models, including TMDD, with execution and consistency checks [25]. Kratos contributes an integrated workflow for calibrating and estimating a fixed, approved full TMDD model: schema-validated actions connect bounded multistart calibration to nlmixr2 FOCEi, diagnostics and reporting, while saved attempts, numerical cross-checks and explicit failure flags make the analysis inspectable. The contribution therefore centres on the integrated estimation workflow and its verification. No head-to-head comparison with PKGPT or GRASP was performed, so superiority over those systems is not established. The reported evidence concerns the deterministic rules-based implementation; compatibility with an optional LLM planner is an architectural feature, not a demonstrated LLM capability. This design supports traceability expectations in pharmacometric analysis [14, 23]. Storing figure data, seeds and hashes alongside each result also supports FAIR reuse of analyses [44].

The planner results argue for caution. In the default setting, the arms were indistinguishable by construction, because the stopping rule was met before any adaptive proposal was made. In the stress setting, the adaptive planner was exercised for a single round and reached the same optima as Latin hypercube multistart with the same number of attempts. Two explanations are likely. First, a bounded trust-region optimiser started from a space-filling design is already a strong baseline for a low-dimensional, smooth objective. Second, the adaptive heuristics are designed for failure modes—boundary solutions and numerically failing regions—that were rare here. The value of the agentic layer in this version therefore lies in orchestration, validation and record keeping rather than in faster optimisation, and the claim of reduced analyst effort rests on the elimination of manual steps rather than on a measured comparison with analysts.

This study has limitations:

- **Aggregate data.** The mAb data are published mean profiles. The fits are therefore pooled, cross-molecule descriptions of a “generic” antibody rather than population PK analyses of any product, and their variance components and standard errors must not be interpreted at the patient level.
- **Data harmonisation and source metadata.** The dataset provenance and digitisation methods are documented in the companion manuscript [36], but Kratos used configuration-declared input units and a common molecular weight, free-drug observation model, bolus approximation for IV doses and drug-naive target baseline. These settings should not be interpreted as verified molecule- or study-specific assay, molecular-weight, infusion-duration or target metadata.

- **VPC.** The VPC was not prediction-corrected, and it indicated that the spread of the pooled data was not reproduced.
- **Censored data.** Censored observations are currently rejected rather than modelled.
- **Limited recovery study.** Parameter recovery was evaluated on a single simulated dataset with a small bootstrap, not on a replicated simulation–estimation study, because each FOCEi fit took tens of minutes on the workstation used.
- **Planner evaluation.** The empirical comparison was restricted to rules-based calibration. The optional LLM interface was tested only with mocked responses; no live LLM was evaluated for proposal quality, recovery from failures, reproducibility or added benefit.
- **Analyst effort.** No study measured analyst time against a manual workflow.

Future work will add censored-data likelihoods, prediction-corrected VPCs, versioned TMDD approximations selectable through recorded decisions, replicated simulation–estimation studies, a prospective comparison with manual analyses and a separately specified evaluation of a live LLM planner against the rules-based baseline.

Conclusions

This proof-of-concept evaluation shows that a constrained, auditable agentic workflow using deterministic rules-based planning can carry TMDD analyses from raw data to population estimates and reports without analyst intervention, recover known parameters, and keep numerical failures visible and traceable. In this evaluation, its benefit came from automation, validation and provenance rather than from faster optimisation. Applied to pooled human mAb mean profiles, it produced consistent, transferable linear PK and absorption parameters and correctly signalled that target-mediated parameters were not identifiable from these data. The optional LLM planner remains unevaluated, and the findings do not establish LLM-driven modelling performance.

Declarations

Funding. The author received no specific funding for this work.

Competing interests. The author declares no competing interests.

Ethics approval and consent. Not applicable. The study used only simulated data and aggregate, previously published concentration–time profiles; no human participants were enrolled.

Data, material and code availability. The Kratos source code, configurations, the molecule-grouped split manifest, the simulated datasets with their generating parameters, and the scripts that regenerate every number, table and figure in this manuscript from saved run artifacts are available at

https://github.com/devamdes/Kratos-An-Auditable-Agentic-Workflow-for-TMDD-Model-Calibration-and-Population-Estimation. The mAb mean-profile dataset is the harmonised, digitised clinical PK database described by Desai and Cristofoletti [36]. The underlying published profiles and source-publication mapping are identified in Table 1 and the references of that manuscript. The compiled dataset is available from the corresponding author of the source study, Rodrigo Cristofoletti (rcristofoletti@ufl.edu), upon reasonable request, as stated in its data-availability declaration. The molecule-grouped Part A/Part B allocation used for Kratos is described in Online Resource 1.

Authors contributions. D.A.D. and R.C conceived the study, D.A.D designed and implemented the software, performed the analyses and wrote the manuscript.

References

1. Mager DE, Jusko WJ (2001) General pharmacokinetic model for drugs exhibiting target-mediated drug disposition. J Pharmacokinet Pharmacodyn 28:507–532. https://doi.org/10.1023/A:1014414520282

2. Lobo ED, Hansen RJ, Balthasar JP (2004) Antibody pharmacokinetics and pharmacodynamics. J Pharm Sci 93:2645–2668. https://doi.org/10.1002/jps.20178

3. Wang W, Wang EQ, Balthasar JP (2008) Monoclonal antibody pharmacokinetics and pharmacodynamics. Clin Pharmacol Ther 84:548–558. https://doi.org/10.1038/clpt.2008.170

4. Dirks NL, Meibohm B (2010) Population pharmacokinetics of therapeutic monoclonal antibodies. Clin Pharmacokinet 49:633–659. https://doi.org/10.2165/11535960-000000000-00000

5. Ryman JT, Meibohm B (2017) Pharmacokinetics of monoclonal antibodies. CPT Pharmacometrics Syst Pharmacol 6:576–588. https://doi.org/10.1002/psp4.12224

6. Mager DE, Krzyzanski W (2005) Quasi-equilibrium pharmacokinetic model for drugs exhibiting target-mediated drug disposition. Pharm Res 22:1589–1596. https://doi.org/10.1007/s11095-005-6650-0

7. Gibiansky L, Gibiansky E, Kakkar T, Ma P (2008) Approximations of the target-mediated drug disposition model and identifiability of model parameters. J Pharmacokinet Pharmacodyn 35:573–591. https://doi.org/10.1007/s10928-008-9102-8

8. Dua P, Hawkins E, van der Graaf PH (2015) A tutorial on target-mediated drug disposition (TMDD) models. CPT Pharmacometrics Syst Pharmacol 4:324–337. https://doi.org/10.1002/psp4.41

9. Peletier LA, Gabrielsson J (2012) Dynamics of target-mediated drug disposition: Characteristic profiles and parameter identification. J Pharmacokinet Pharmacodyn 39:429–451. https://doi.org/10.1007/s10928-012-9260-6

10. Karlsson MO, Savic RM (2007) Diagnosing model diagnostics. Clin Pharmacol Ther 82:17–20. https://doi.org/10.1038/sj.clpt.6100241

11. Savic RM, Karlsson MO (2009) Importance of shrinkage in empirical Bayes estimates for diagnostics: Problems and solutions. AAPS J 11:558–569. https://doi.org/10.1208/s12248-009-9133-0

12. Nguyen THT, Mouksassi M-S, Holford N, et al (2017) Model evaluation of continuous data pharmacometric models: Metrics and graphics. CPT Pharmacometrics Syst Pharmacol 6:87–109. https://doi.org/10.1002/psp4.12161

13. Byon W, Smith MK, Chan P, et al (2013) Establishing best practices and guidance in population modeling: An experience with an internal population pharmacokinetic analysis guidance. CPT Pharmacometrics Syst Pharmacol 2:e51. https://doi.org/10.1038/psp.2013.26

14. Bonate PL, Strougo A, Desai A, et al (2012) Guidelines for the quality control of population pharmacokinetic–pharmacodynamic analyses: An industry perspective. AAPS J 14:749–758. https://doi.org/10.1208/s12248-012-9387-9

15. Bies RR, Muldoon MF, Pollock BG, et al (2006) A genetic algorithm-based, hybrid machine learning approach to model selection. J Pharmacokinet Pharmacodyn 33:195–221. https://doi.org/10.1007/s10928-006-9004-6

16. Li X, Sale M, Nieforth K, et al (2024) pyDarwin: A machine learning enhanced automated nonlinear mixed-effect model selection toolbox. Clin Pharmacol Ther 115:758–773. https://doi.org/10.1002/cpt.3114

17. Li X, Sale M, Nieforth K, et al (2024) pyDarwin machine learning algorithms application and comparison in nonlinear mixed-effect model selection and optimization. J Pharmacokinet Pharmacodyn 51:785–796. https://doi.org/10.1007/s10928-024-09932-9

18. Chen X, Nordgren R, Belin S, et al (2024) A fully automatic tool for development of population pharmacokinetic models. CPT Pharmacometrics Syst Pharmacol 13:1784–1797. https://doi.org/10.1002/psp4.13222

19. Yao S, Zhao J, Yu D, et al (2023) ReAct: Synergizing reasoning and acting in language models. In: International conference on learning representations (ICLR)

20. Cloesmeijer ME, Janssen A, Koopman SF, et al (2024) ChatGPT in pharmacometrics? Potential opportunities and limitations. Br J Clin Pharmacol 90:360–365. https://doi.org/10.1111/bcp.15895

21. Shin E, Yu Y, Bies RR, Ramanathan M (2024) Evaluation of ChatGPT and Gemini large language models for pharmacometrics with NONMEM. J Pharmacokinet Pharmacodyn 51:187–197. https://doi.org/10.1007/s10928-024-09921-y

22. Cha HJ, Choe K, Shin E, et al (2025) Leveraging large language models in pharmacometrics: Evaluation of NONMEM output interpretation and simulation capabilities. J Pharmacokinet Pharmacodyn 52:34. https://doi.org/10.1007/s10928-025-09982-7

23. Shahin MH, Goswami S, Lobentanzer S, Corrigan BW (2025) Agents for change: Artificial intelligent workflows for quantitative clinical pharmacology and translational sciences. Clin Transl Sci 18:e70188. https://doi.org/10.1111/cts.70188

24. Kwack H, Kong H, Lim J, et al (2026) PKGPT: Expert-orchestrated recursive LLM agent for automated NONMEM PopPK modeling with human benchmarking. Pharmaceutics 18:501. https://doi.org/10.3390/pharmaceutics18040501

25. Bazgir O, Manthapuri V, Rattsev I, Jafarnejad M (2025) GRASP: Graph Reasoning Agents for Systems Pharmacology with Human-in-the-Loop. arXiv [preprint]. https://doi.org/10.48550/arXiv.2512.05502

26. Branch MA, Coleman TF, Li Y (1999) A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. SIAM J Sci Comput 21:1–23. https://doi.org/10.1137/S1064827595289108

27. Virtanen P, Gommers R, Oliphant TE, et al (2020) SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2

28. McKay MD, Beckman RJ, Conover WJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21:239–245. https://doi.org/10.1080/00401706.1979.10489755

29. Fidler M, Wilkins JJ, Hooijmaijers R, et al (2019) Nonlinear mixed-effects model development and simulation using nlmixr and related R open-source packages. CPT Pharmacometrics Syst Pharmacol 8:621–633. https://doi.org/10.1002/psp4.12445

30. Wang W, Hallow KM, James DA (2016) A tutorial on RxODE: Simulating differential equation pharmacometric models in R. CPT Pharmacometrics Syst Pharmacol 5:3–10. https://doi.org/10.1002/psp4.12052

31. Wang Y (2007) Derivation of various NONMEM estimation methods. J Pharmacokinet Pharmacodyn 34:575–593. https://doi.org/10.1007/s10928-007-9060-6

32. Petzold L (1983) Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. SIAM J Sci Stat Comput 4:136–148. https://doi.org/10.1137/0904010

33. Hooker AC, Staatz CE, Karlsson MO (2007) Conditional weighted residuals (CWRES): A model diagnostic for the FOCE method. Pharm Res 24:2187–2197. https://doi.org/10.1007/s11095-007-9361-x

34. Bergstrand M, Hooker AC, Wallin JE, Karlsson MO (2011) Prediction-corrected visual predictive checks for diagnosing nonlinear mixed-effects models. AAPS J 13:143–151. https://doi.org/10.1208/s12248-011-9255-z

35. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York

36. Desai DA, Cristofoletti R (2026) Characterization of inter-antibody variability in clinical pharmacokinetics of monoclonal antibodies using low-dimensional neural ordinary differential equations (NODE). CLINICAL PHARMACOLOGY & THERAPEUTICS 119.

37. Kumar M, Lanke S, Yadav A, et al (2024) Inter-Antibody Variability in the Clinical Pharmacokinetics of Monoclonal Antibodies Characterized Using Population Physiologically Based Pharmacokinetic Modeling. Antibodies 13:54. https://doi.org/10.3390/antib13030054

38. Rohatgi A. WebPlotDigitizer, version 5.2. https://automeris.io

39. Sheiner LB, Beal SL (1981) Some suggestions for measuring predictive performance. J Pharmacokinet Biopharm 9:503–512. https://doi.org/10.1007/BF01060893

40. Harris CR, Millman KJ, Walt SJ van der, et al (2020) Array programming with NumPy. Nature 585:357–362. https://doi.org/10.1038/s41586-020-2649-2

41. R Core Team (2025) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

42. Allaire JJ, Teague C, Xie Y, Dervieux C (2022) Quarto

43. Davda JP, Dodds MG, Gibbs MA, et al (2014) A model-based meta-analysis of monoclonal antibody pharmacokinetics to guide optimal first-in-human study design. mAbs 6:1094–1102. https://doi.org/10.4161/mabs.29095

44. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18

Online Resources

Online Resource 1. Molecule composition of Part A and Part B (online_resource_1_split.csv).

Online Resource 2. Cross-engine agreement by case, automated test-suite summary, complete calibration attempt histories and planner-evaluation details.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- onlineresource1split.csv
- onlineresources.docx